Discovering Music Structure via Similarity Fusion

Meng, Anders

Publication date:
2007

Discovering Music Structure via Similarity Fusion

J. Arenas-García, E. Parrado-Hernández
Dep. Signal Theory and Communications
Universidad Carlos III de Madrid
28911 Leganés, Spain
{jarenas,emipar}@tsc.uc3m.es

A. Meng, L. K. Hansen and J. Larsen
Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Kongens Lyngby, Denmark
{am,lhk,jl}@imm.dtu.dk

Abstract

Automatic methods for music navigation and music recommendation exploit the structure in the music to carry out a meaningful exploration of the "song space". To get a satisfactory performance from such systems, one should incorporate as much information about songs similarity as possible; however, how to do so is not obvious. In this paper, we build on the ideas of the Probabilistic Latent Semantic Analysis (PLSA) that have been successfully used in the document retrieval community. Under this probabilistic framework, any song will be projected into a relatively low dimensional space of "latent semantics", in such a way that all observed similarities can be satisfactorily explained using the latent semantics. Therefore, one can think of these semantics as the real structure in music, in the sense that they can explain the observed similarities among songs. The suitability of the PLSA model for representing music structure is studied in a simplified scenario consisting of 4412 songs and two similarity measures among them. The results suggest that the PLSA model is a useful framework to combine different sources of information, and provides a reasonable space for song representation.

1 Introduction

Given two songs, most people would agree that it is possible to tell if the two songs are similar or not. However, similarity between songs can be "defined" in many different ways: They may have the same beat, the same guitar sound, the same lead singer, etc. One may also extend the domain beyond the sound-based context, and state that two songs are similar if they were produced in the same year or if they are targeted to the same audience. In short, similarities among songs are many and varied.

If we are interested in finding out a model for music structure, we would like to integrate as much information about songs similarity as possible. Looking at the literature about music content-based search and retrieval systems, we can find many different solutions to how the information of the chosen features should be combined in order to build a unique space for song representation. In [1], for instance, some low level features such as the loudness, pitch, brightness, bandwidth and harmonicity, are aggregated by the mean, variance and autocorrelation. In [2], the MFCCs are binned using a vector quantization tree in which the decision thresholds are set to maximize the mutual information between the inputs and the labels of a training set. In other approaches, such as [3, 4], the data cloud of low level features is modeled using a probability distribution, typically estimated using a Gaussian mixture model (GMM). For many low-level features this is a sensible thing to do and well justified given the empirical distribution of the features. But as the feature set is expanded from, say MFCCs or zero crossing rates, to playlist co-occurrence, production year, or blog-gossip, it becomes increasingly unlikely that any practical family of distributions will suffice to model the observations, and thus to build a reasonable similarity space.

In this paper we propose a generalized framework for building representation spaces for songs using a combination of different (possibly redundant) sources of information regarding song similarity. Our approach makes use of the ideas of Probabilistic Latent Semantic Analysis (PLSA) [5, 6], which have been successfully applied in web document retrieval, including the possibility of combining heterogeneous similarity measures between documents, such as the appearance of common words or common links [6]. The basic idea is to project the songs into a space of relatively small dimension (the latent semantics) in such a way that all observed similarities can be satisfactorily explained using the latent semantics. In this way, the "overall" distance between two songs can be determined from the latent semantics only; in this sense, we can say that the semantics provide a meaningful representation for music structure. Furthermore, as in the document retrieval case, the application of PLSA simplifies the implementation of music recommendation systems, significantly reducing the computational burden of the song retrieval phase.

This analogy between songs and documents can be regarded as a purely technical convenience, but might also start a new line of thinking in which songs aspects (e.g. timbre, frequency representation, etc.) are interpreted as "words". In any case, if this is a fruitful analogy, future research could build models for music using the elaborated machinery already deployed for web-mining.

The rest of the paper is organized as follows: Section 2 introduces the different levels of representation for music analysis that will be used throughout the paper, while Section 3 reviews the formulation for the Generalized version of PLSA that can be used for building models for music structure that simultaneously exploit the information from multiple measures of similarity between songs. Different algorithms can be used to adjust the parameters of the PLSA model; in this paper we consider Non-negative Matrix Factorization (NMF) algorithms as described in Section 4. In Section 5 we evaluate the possibilities of the approach by carrying out experiments in a simplified scenario, and in Section 6 we extract some conclusions about the work, and discuss lines for future research.

2 Music Representation Levels

In this section, we introduce some notation, and define the different levels for music representation that will be considered along the paper:

- *Songs*: This level corresponds to a collection of pieces of music. The set of all songs will be denoted as $\{s_l\}_{l=1}^{L}$.
- *Similarities*: Each of the different criteria that we use to measure distances between songs. In this paper, we will consider that each similarity criterion is characterized by a set of clusters (e.g., $c_j^{(k)}$ for the jth group of the kth similarity), and that each song s_l is defined by a certain distribution over the clusters of each similarity criterion, subject to restrictions:

  $$P(c_j^{(k)}|s_l) > 0, \ \ \forall j,k \qquad \text{and} \qquad \sum_{j=1}^{n_k} P(c_j^{(k)}|s_l) = 1, \ \ \forall k$$

  where n_k is the number of groups along the kth similarity dimension.

  In a real situation, there are different possibilities to estimate this similarity information. For instance, when song recordings are available, we can extract "sound features" from the music (e.g., zero crossing rate, MFCCs, etc) and carry out a hard or soft clustering in the resulting feature space. Sometimes, we also have access to metadata (e.g. music genre) that can be interpreted as the labels of a multi-class problem. In such cases, we can either use the class membership of each song or, alternatively, the outputs of a classification system operating on the "sound features" to predict this information. Other sources of information (e.g., playlist co-occurance) can also be exploited.
- *Latent Semantics*: This is the representation space where songs are projected to get a useful representation. As with songs, each semantic, $z_i, \ i = 1, \ldots N$, is represented by a certain distribution along each similarity dimension. These latent semantics are not known a priori, but have to be determined from the set of songs and their representations along the different similarity criteria. However, once the semantics space is built, it provides a good overall representation for music structure, in the sense that similarities among songs, according to all considered criteria, can be reasonably explained from the semantics.

3 Generalized PLSA for music similarities fusion

Our model for music structure is based on the Probabilistic Latent Semantics Analysis (PLSA) that has been successfully used in the analysis and retrieval of text documents [5]. The analogy is as follows: songs (documents) can belong to a set of hidden and unknown groups, $\{z_i\}_{i=1}^{N}$, i.e., the latent semantics. We assume soft membership, so that each song can be represented as a distribution over the different hidden states, thus satisfying the constraint:

$$\sum_{i=1}^{N} P(z_i|s_l) = 1 \tag{1}$$

where $P(z_i|s_l)$ is the probability that song s_l belongs to the semantic group z_i. Semantics give a good insight into music structure: Two songs that belong (to a large degree) to the same semantic are assumed to be similar, both with respect to the semantics, and with respect to all similarity criteria.

Next, each (hidden) group of songs is characterized by some cluster distribution over each of the similarity dimensions we are considering, i.e.,

$$z_i : P(c_1^{(k)}|z_i), P(c_2^{(k)}|z_i), \ldots, P(c_{n_k}^{(k)}|z_i)$$

Of course, each of these distributions have to be a real distribution, i.e.,

$$P(c_j^{(k)}|z_i) > 0, \;\; \forall j,k \qquad \text{and} \qquad \sum_{j=1}^{N_k} P(c_j^{(k)}|z_i) = 1, \forall k \tag{2}$$

Now, we can express $P(c_j^{(k)}|s_l)$ through the expansion

$$P(c_j^{(k)}|s_l) = \sum_{i=1}^{N} P(c_j^{(k)}|z_i, s_l)P(z_i|s_l) = \sum_{i=1}^{N} P(c_j^{(k)}|z_i)P(z_i|s_l) \tag{3}$$

where we are assuming that all the knowledge about the cluster distribution is propagated via the semantic groups.

As it is usual in the PLSA approach, we assume that $P(c_j^{(k)}|s_l)$ are unknown, but we have access to some estimations of these quantities that we will denote as $\tilde{P}(c_j^{(k)}|s_l)$. Then, for each similarity criterion, we would like to find the set of probabilities $P(c_j^{(k)}|z_i)$ and $P(z_i|s_l)$ that maximize the likelihood of our observations,

$$\prod_{j,l} P(c_j^{(k)}|s_l)^{\tilde{P}(c_j^{(k)}|s_l)}$$

Finally, taking logarithms, and introducing the decomposition model for $P(c_j^{(k)}|s_l)$ [Eq. (3)], we get the following set of log-likelihoods to be maximized:

$$L_k = \sum_{j,l} \tilde{P}(c_j^{(k)}|s_l) \log \sum_{i=1}^{N} P(c_j^{(k)}|z_i)P(z_i|s_l), \tag{4}$$

for $k = 1, \ldots, K$, K being the total number of available similarities.

Note that the different log-likelihoods for different similarities cannot be maximized independently since they are coupled through terms $P(z_i|s_l)$. As in [6], we propose to maximize the following combined log-likelihood function

$$L = \sum_{k=1}^{K} \alpha_k L_k \tag{5}$$

where α_k, satisfying $\sum_k \alpha_k = 1$, measures the importance assigned to the kth similarity. Note that, proceeding this way, we can adjust models for music that are specially good at explaining different similarities (for instance, we can obtain a model which is specially good at explaining similarity in the co-play dimension, while still integrating some of the information in the other similarity dimensions). The maximization of this mixed log-likelihood w.r.t. $P(c_j^{(k)}|z_i)$ and $P(z_i|s_l)$ can be carried out using different methods, such as versions of the Expectation-Maximization algorithm, or the Non-negative Matrix Factorization (NMF) approach discussed in Section 4.

Song retrieval procedure

Apart from providing a meaningful approach for discovering music structure, semantics are also useful from other points of view. For instance, in this subsection we analyze their application to song recommendation systems. Once the PLSA model has been trained, we can use the latent semantics for song retrieval using very compact expressions. In this way, the probability that any song in the dataset should be recommended given some query song, s_q, can be calculated using

$$\begin{aligned} P(s|s_q) &= \sum_{i=1}^{N} P(s|z_i, s_q)P(z_i|s_q) = \sum_{i=1}^{N} P(s|z_i)P(z_i|s_q) \\ &= \sum_{i=1}^{N} \frac{P(z_i|s)P(s)}{P(z_i)} P(z_i|s_q) \end{aligned} \tag{6}$$

where we have used the assumption that song probability distributions propagate through the latent semantics in replacing $P(s|z_i, s_q)$ by $P(s|z_i)$, and where $P(s)$ is the a priori probability of each song, that can be estimated, e.g., using a measure of song popularity. Finally, the a priori probabilities assigned to each latent semantic can be precalculated using

$$P(z_i) = \sum_{l} P(z_i|s_l)P(s_l), \quad i = 1, \ldots, N \tag{7}$$

Note that the complexity in evaluating (6) grows linearly with the number of latent semantics. This is a very important advantage with respect to the case in which similarity clusters were considered directly. Effectively, if the expansion were made with respect to all clusters in all similarities, we would get

$$P(s|s_q) = \sum_{j_1 \cdots j_K} P(s|c_{j_1}^{(1)}, \ldots, c_{j_K}^{(K)})P(c_{j_1}^{(1)}, \ldots, c_{j_K}^{(K)}|s_q) \tag{8}$$

In this sense, we can interpret the PLSA model as a bottleneck that is reducing the complexity of the problem from all possible combinations of clusters ($\prod_i n_{c_i}$) to just the number of hidden states (N). Nevertheless, maximization of combined likelihood (5) assures that the latent semantics retain as much information as possible about the different similarity dimensions that are taken into account.

Furthermore, the fact that PLSA is a probabilistic framework provides a lot of flexibility when carrying out other search tasks, such as constraining the search to songs that belong to a certain cluster (i.e., calculating $P(s|s_q, c_j^{(k)})$), or for incorporating user preferences into the model (what can be done, for instance, by allowing the user to tune parameters α_k).

4 NMF optimization of the PLSA model

In [7] the authors showed the relation between Non-negative Matrix Factorization (NMF) using Kullback-Leibler divergence and PLSA. In this section, we propose a multiplicative NMF update scheme for determining the unknown parameters of the combined PLSA model. Instead of minimizing the log-likelihood cost function (5), we will solve the following NMF optimization problem

$$\min_{\mathbf{W}^{(k)}, \mathbf{H}} \quad \sum_{k=1}^{K} \alpha_k ||\tilde{\mathbf{P}}^{(k)} - \mathbf{W}^{(k)}\mathbf{H}||_F^2, \quad \text{subject to } \mathbf{W}^{(k)} \geq \mathbf{0}, \mathbf{H} \geq \mathbf{0} \tag{9}$$

where $||\mathbf{A}||_F^2$ denotes the squared Frobenius norm of a matrix, hence $\sum_{i,j} \mathbf{A}_{i,j}^2$, and $\mathbf{A} \geq 0$ means that all elements in $\mathbf{A}$ are non-negative.

By proper normalization of $\mathbf{W}^{(k)}$ and $\mathbf{H}$ we can ensure the validity of the following interpretation

$$\left(\mathbf{W}^{(k)}\mathbf{H}\right)_{j,l} = \sum_{i=1}^{N} P(c_j^{(k)}|z_i)P(z_i|s_l), \tag{10}$$

from which, $\mathbf{W}_{j,i}^{(k)} = P(c_j^{(k)}|z_i)$ and $\mathbf{H}_{i,l} = P(z_i|s_l)$.

One way of minimizing (9) is to use a multiplicative update method, see e.g. [8]. Assuming the algorithm has converged to some point within the feasible region where $\mathbf{W}^{(k)} > \mathbf{0}$ and $\mathbf{H} > \mathbf{0}$, it can be shown that this point is a stationary point, which may or may not be a local minimum (see [8] for a more complete discussion about algorithms for solving NMF types of problems).

The following pseudo-code provides a multiplicative update scheme for solving the NMF problem given in (9). It can be easily seen that, if matrices $\mathbf{W}^{(k)}$ and $\mathbf{H}$ are initialized to strictly positive values, then these matrices remain positive throughout the iterations, as a consequence of the multiplicative update scheme.

1. Initialize $\mathbf{W}^{(k)}$ and $\mathbf{H}$.
2. Iterate:
   (a)
   $$\mathbf{W}^{(k)}_{j,i} = \frac{(\tilde{\mathbf{P}}^{(k)}\mathbf{H}^T)_{j,i}}{(\mathbf{W}^{(k)}\mathbf{H}\mathbf{H}^T)_{j,i} + 10^{-9}}\mathbf{W}^{(k)}_{j,i}, \quad \text{for } k = 1, \ldots, K.$$
   (b) Normalize $\mathbf{W}^{(k)}$ such that $\sum_j \mathbf{W}^{(k)}_{j,i} = 1$, for $i = 1, \ldots, N$ and $k = 1, \ldots, K$
   (c)
   $$\mathbf{H}_{i,l} = \frac{\sum_k \alpha_k \left(\mathbf{W}^{(k)}\tilde{\mathbf{P}}^{(k)}\right)_{i,l}}{\sum_k \alpha_k \left(\mathbf{W}^{(k)^T}\mathbf{W}^{(k)}\mathbf{H}\right)_{i,l} + 10^{-9}}\mathbf{H}_{i,l}$$
3. Repeat 2 until some convergence criteria is met.

5 Experiments

5.1 Dataset description

To illustrate the suitability of the PLSA model we have used a data set which was downloaded from the online music site `http://www.garageband.com/`[1]. The music site has an online music reviewing system, which allows artists to review uploaded music. The reviews can provide valuable information to the artist on his/her performance, use of instrumentation, etc. The complete data set consists of 4412 song titles (MP3 files), with their corresponding music reviews and genre labels. A first analysis of the data set revealed that each song was reviewed on the average by 80 people, where each review consists of ≈ 71 words. In the following, whenever we write "song review" we will refer to all reviews for a particular song.

The genre taxonomy at `www.garageband.com` has a flat structure and has been designed from artist similarity[2]. In total, 47 genres are used. However, in order to minimize confusion among genres, we chose to fuse the taxonomy into 18 categories. The resulting 18-genre-taxonomy information is summarized in Table 1.

5.2 Song similarity extraction

In this paper we consider two sources of similarity among songs: the first of them exploits the similarity among the "textual features" extracted from the reviews of each song title, while the second is based on the available genre information.

For the reviews, we can have direct access to $\tilde{P}(c_j^{(k)}, s)$, which could be given by the term frequencies. Due to the high-dimensionality of the textual features, however, we preferred to extract context information from the textual features using clustering. Regarding the genre similarity, we extract first some "sound features" from the audio, and use them as inputs to a genre classifier, whose probabilistic predictions can be used as the similarity information.

[1] Downloaded in November, 2005.

[2] For instance, bands describing the genre "alternative pop" are: LIVE, REM and Sheryl Crow.

GENRE	# TITLES	GENRE	# TITLES	GENRE	# TITLES	GENRE	# TITLES
Acoustic	90	Electronica	388	Punk	493	Spoken Word	176
Alternative Rock	840	Folk	289	R&B	39	Techno	260
Blues	127	Hard Rock	450	Rap	84	World	102
Classical	73	Jazz	89	Reggae	118		
Country	100	Pop	126	Rock	568		

Table 1: Distribution of songs among genres.

Similarity $c^{(1)}$ contexts in reviews (unsupervised): To process the textual information we applied the bag-of-words approach,using the "RainBow" toolkit [9] for preprocessing the song reviews. This program efficiently extracts term-document matrices from collections of text data. In the preprocessing step, words which are shorter than two letters (after stemming) and those which did not occur in at least 10 song reviews were pruned away. The raw term-frequency counts were used to calculate a normalized TFIDF (term-frequency inverse document frequency) matrix of dimension 31560×4412. From the normalized TFIDF matrix, 90 contexts were extracted by using a clustering algorithm based on NMF. A simple Alternating Least Squares(ALS) strategy was used in an iterative manner to extract the context information $[P(w_i|c_j^{(2)})]$ and song membership $[P(c_j^{(2)}|s)]$. At this stage, 90 components seemed a resonable number, in terms of squared error, to provide a good representation of the TFIDF matrix.

Similarity $c^{(2)}$ music genre (supervised): Although genre labels could be straightforwardly use as the similarity features [hence, $\tilde{P}(c_j^{(k)}|s_l)$ would be either 0 or 1], we followed a potentially more useful approach, based on training a clasifier that predicts the a posteriori probability of each of the genres for any song. In this way, we allow for a smoother gradation in this similarity dimension.

The inputs to the classification scheme are a set of "sound features" which are the coefficients of a multivariate autoregressive (MAR) model [10]: Basically, we extract Mel Frequency Cepstral Coefficients (MFCC) using a window length of 20 msec., and compose a time series with the first 6 MFCCs (excluding the first one which is associated to volume) over a time frame of 1 sec. For each such block, we adjust an MAR model of lag three: $\mathbf{x}_n = \sum_{p=1}^{3} \mathbf{A}_p \mathbf{x}_{n-p} + \mathbf{e}_p$, where $\mathbf{x}_n$ is used to denote a vector of MFCC features inside the window. The values of matrices $\mathbf{A}_p$, $p = 1, 2, 3$, together with the mean and covariance of the residuals, $\mathbf{e}_n$, are concantenated into a single feature vector (MAR feature) of length 135. Using 30 s. for each title in the data set, each song can be represented by 30 MAR vectors.

Finally, we trained a neural network taking the MAR features as inputs, and the genre information as the labels. The neural network consists of a non-linear feature extraction phase, using the rKOPLS algorithm of [11], followed by a linear classifier. Though each MAR feature was assigned to just one genre, soft membership of the music snippets to the different genres was determined using late fusion. Hence, simply counting how many of the MAR features of each song were classified into each of the genres.

5.3 Results and discussion

The modified NMF algorithm suggested in Section 4 was run with a varying dimension of the "latent semantics" ranging from 3 to 50. We have also considered different values of α[3] between 0 (only similarity $c^{(2)}$ was used) and 1 (only $c^{(1)}$ was used). Experimental results show that going much further than 40 semantics does not improve the results significantly. Note that this number is much smaller than the number of possible combinations using one cluster from each similarity criterion, and thus the PLSA approach provides a much more compact and convenient representation for song recommendation than the direct use of (8). In each run of the NMF algorithm, the algorithm was stopped after 1000 iterations, where convergence was found to be complete in all cases.

Left side of Figure 1 shows the distance between the empirical distribution $\tilde{P}(c_j^{(1)}|s_l)$ and the PLSA model, i.e., the value L_1 given by (4), as a function of varying α. With $\alpha = 0$, the latent space is estimated purely from the $c^{(2)}$ similarity measure, which explains the high approximation error

[3]Since we use two similarities, we denote α_1 by α. Therefore $\alpha_2 = 1 - \alpha$.

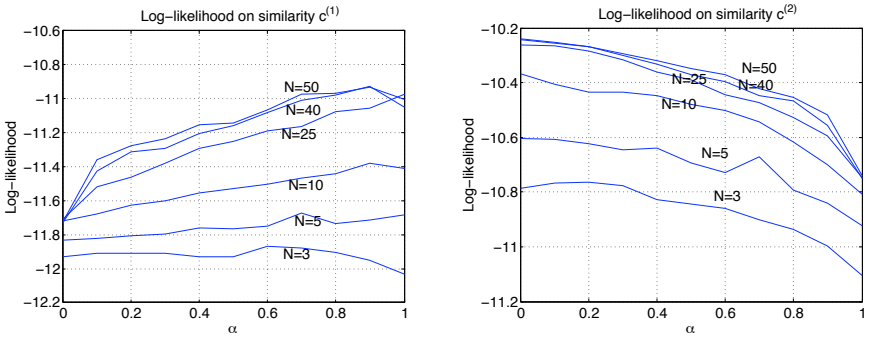

Figure 1: Log-likelihood of the PLSA model with respect to reviews (left) and genres (right).

Semantic	Z1 (0.08)	Z2 (0.06)	Z3 (0.06)	Z4 (0.05)	Z5 (0.05)	Z6 (0.04)	Z7 (0.04)	Z8 (0.04)	Z9 (0.04)
Genres	punk	techno	acoustic	jazz	electronica	pop	country	classical	reggae
Words	punk	dance	acoustic	jazz	samples	beatles	country	piano	reggae
	ska	trance	mandoline	piano	dance	chorus	female	classic	ska
	horns	sequencing	folk	sax	ambient	pop	fiddle	strings	horns
	blink	synth	cello	jazzy	synth	harmonies	harmonica	ambient	funk
	emo	techno	female	garageband	electronic	goo	steel	piece	marley

Table 2: Nine most significant semantics for $N = 50$ and $\alpha = 0$. Numbers in brackets by the semantics are their prior probability $P(z_i)$. All probabilities of the dominant genre given the semantic $P(c_k^{(2)}|z_i)$ were above 0.93 (indeed most of them lie above 0.99).

to the real distribution. Conversely, when $\alpha = 1$ (this corresponds to considering only the $c^{(1)}$ similarity) a much better solution, with respect to similarity $c^{(1)}$, is obtained. It is interesting to notice (see also [6]) how for $N \geq 40$ the log-likelihood is larger in the range $\alpha = 0.6$ to $\alpha = 0.9$ than for $\alpha = 1$. In other words, incorporating some information about the $c^{(2)}$ similarity, serves to improve the capabilities of the PLSA model to represent similarity in dimension $c^{(1)}$.

Right side of Figure 1 shows the log-likelihood of the PLSA model for similarity $c^{(2)}$. In view of these results, one can conclude that using only one of the similarity dimensions results in a very small likelihood of the observations associated to the other similarity. However, there is a wide range $\alpha \in [0.4, 0.7]$ for which the semantics simultaneously offer a good representation of both similarities.

We can get deeper insight about the PLSA model by looking at the distributions $P(c^{(1)}|z_i)$ and $P(c^{(2)}|z_i)$. For instance, for $\alpha = 0$, semantics are pretty much aligned with the genres: there is a dominant genre for every semantic (see Table 1). In addition, note how the most relevant words (coming from the subjective reviews of the songs) for each semantic seem to be connected to the description of the corresponding genre. This fact reveals that there is some underlying structure in the data set.

This is even more clear when looking at the semantics displayed in Table 3. The introduction of the textual similarity merges close genres into the same semantic. This is the case in semantics Z2 and Z7. It is also interesting to see the descriptions of semantic Z3 in terms of relevant words: here words like “ambient” and “relaxing” explain the merging of techno, electronica and classical music, as opposed to the words describing a pure electronica semantic as the one in Table 2. All in all, we can see that the combined use of both similarities enables us to discover sensible groups of music pieces beyond the isolated information provided by each similarity on its own.

6 Conclusions

In this paper we have presented an extension of the PLSA framework for its application in music. Basically, the proposed PLSA model works by projecting the songs into a latent semantic space. This space is obtained by maximizing a combined log-likelihood which takes into account different

Semantic	Z2 (0.05)	Z3 (0.04)	Z4 (0.04)	Z7 (0.04)	Z9 (0.03)
Genres	acoustic (0.790) folk (0.133)	techno (0.299) electronica (0.286) classical (0.146)	country (0.290) acoustic (0.198) folk (0.161)	rap (0.746) spoken word (0.187)	rap (0.519) r&b (0.225) reggae (0.177)
Words	acoustic folk beautiful accoustic mandolin	ambient relaxing electronic pads chill	country fiddle slide steel folk	funny spoken comedy poetry word	funk funky horns wah horn

Table 3: Among the nine most significant semantics for $N = 50$ and $\alpha = 0.5$, there are four "pure" semantics (one dominant genre) and five "mixed" ones (several dominant genres), that we reproduce in this table. Numbers in brackets by the semantics are their prior probabilty while numbers in brackets by the genres are the conditional probability of the genre given the semantic.

sources of similarity between songs. By doing so, the latent semantics can satisfactorily explain all observed similarities and provide a very convenient representation for music structure.

Although more work is needed to study the impact of the PLSA approach on music organization tasks, we think that the analogy between documents and songs promises to be very fruitful, and opens new lines for investigating music structure using the elaborated machinery already deployed for web-mining, and for improving the performance of music recommendation systems.

Acknowledgments

This work has been partly by Spanish Ministry of Education and Science grant CICYT TEC-2005-00992, by Madrid Community grant S-505/TIC/0223 and by the Danish Technical Research Council, through the framework project 'Intelligent Sound', `www.intelligentsound.org` (STVF No. 26-04-0092).

References

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, pp. 27–36, 1996.

[2] J. Foote, "Content-based retrieval of music and audio," in *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, vol. 3229, pp. 138–147, 1997.

[3] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman, "A large scale evaluation of acoustic and subjective music similarity measures," in *Proc. of the Intl. Symp. on Music Information Retrieval*, 2003.

[4] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?," in *Proc. of the Intl. Symp. on Music Information Retrieval*, 2002.

[5] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, pp. 289–296, 1999.

[6] D. Cohn and T. Hofmann, "The Missing Link – A Probabilistic Model of Document Content and Hypertext Connectivity," in *Neural Information Processing Systems 13*, 2001.

[7] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in *SIGIR*, pp. 601–602, 2005.

[8] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca and R. J. Plemmons "Algorithms and Applications for Approximate Nonnegative Matrix Factorization," in *Computational Statistics and Data Analysis. Elsevier. To appear* , 2007.

[9] A. K. McCallum "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," `http://www.cs.cmu.edu/~mccallum/bow`, 1996.

[10] A. Meng and P. Ahrendt and J. Larsen and L. K. Hansen "Temporal Feature Integration for Music Genre Classification," in *IEEE Trans. on Audio, Speech and Language Process.*, 2007.

[11] J. Arenas-García, K. B. Petersen and L. K. Hansen, "Sparse Kernel Orthonormalized PLS for feature extraction in large data sets,", in *Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, MA*, 2007